Title
Folding very short peptides using molecular dynamics.

Journal
PLoS computational biology, 2(4)

ISSN
1553-734X

Authors
Ho, Bosco K
Dill, Ken A

Publication Date
2006-04-01

DOI
10.1371/journal.pcbi.0020027

Peer reviewed

OPEN ACCESS Freely available online

Folding Very Short Peptides Using Molecular Dynamics

Bosco K. Ho*, Ken A. Dill

Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America

Peptides often have conformational preferences. We simulated 133 peptide 8-mer fragments from six different proteins, sampled by replica-exchange molecular dynamics using Amber7 with a GB/SA (generalized-Born/solvent-accessible electrostatic approximation to water) implicit solvent. We found that 85 of the peptides have no preferred structure, while 48 of them converge to a preferred structure. In 85% of the converged cases (41 peptides), the structures found by the simulations bear some resemblance to their native structures, based on a coarse-grained backbone description. In particular, all seven of the β hairpins in the native structures contain a fragment in the turn that is highly structured. In the eight cases where the bioinformatics-based I-sites library picks out native-like structures, the present simulations are largely in agreement. Such physics-based modeling may be useful for identifying early nuclei in folding kinetics and for assisting in protein-structure prediction methods that utilize the assembly of peptide fragments.

Citation: Ho BK, Dill KA (2006) Folding very short peptides using molecular dynamics. PLoS Comput Biol 2(4): e27. DOI: 10.1371/journal.pcbi.0020027

Introduction

Peptide fragments of proteins often have intrinsic propensities for the formation of their native conformations. For example, NMR experiments [1] show that long peptide fragments have native-like conformations [2–7]. Some short peptides in solution have also been shown to adopt their native secondary structures: α helices [8,9] and β hairpins [10–14].

As a consequence, peptide conformational propensities that are taken from the protein databank (PDB) [1–17] are now widely used in protein-structure prediction algorithms. A popular set of peptide fragment conformations is the I-sites library of David Baker and his co-workers [18,19]. Extensive libraries of peptide fragments have now been compiled [20–22] and have become essential elements in protein-prediction methods [23]. From the recent CASP protein–structure prediction competition, it was noted that most of the successful de novo methods use a fragment-based approach [23,24]. Typically, a candidate protein native structure is spliced together from fragments that are extracted from a database of conformations, and then treated to conformational scoring and optimization.

Can physical models capture these conformational propensities of peptides? There is good evidence that they can. First, simple physical models can reproduce the structural biases of certain peptide fragments [25–28]. To date, however, such studies have largely focused on selected peptides that are expected to fold. Our interest here is to know whether physical models can also discriminate peptides that fold from peptides that do not. Second, in molecular dynamics simulations of small peptides, the ensemble of conformers divides into well-defined clusters. This has been found for a penta–β peptide in explicit water [29,30], and for a small α-helical peptide [31]. Third, molecular dynamic simulations of small peptides reproduce the α-helical propensities of certain fragments from the I-sites sequence-structure library [32]. Many models of protein folding kinetics assume that peptide fragments of the chain that have preferred conformations are responsible for nucleating the folding process [33–35].

Here, we study 133 peptide 8-mer fragments from six different proteins of different folds, using replica-exchange molecular dynamics sampling [36] in Amber7, with the parm96 parameters and the GB/SA (generalized-Born/solvent-accessible electrostatic approximation to water) implicit solvent model of Tsui and Case [37]. We chose this force field as it is the only implicit-solvation model that can adequately reproduce the native state of the β hairpin of protein G [38].

We are interested in whether this physical model can identify native-like secondary structures in peptide fragments. If so, it indicates the importance of local interactions in those cases. Our study involves complete coverage of those proteins. For each protein, we systematically generate a series of 8-mer peptide fragments with overlapping sequences from the original protein sequence. Neighboring fragments have a five-residue overlap (and three-residue gap). We chose 8-mers because this length appears adequate to identify elements of structure in PDB studies [19] and because much longer fragments become too expensive for computer simulations. We simulate each peptide using 16 replicas for 5 ns/replica, and keep only the last 1 ns.

In each case, we determine whether the peptide has converged to its native conformation in the folded protein. We consider two measures of convergence. First, we monitor the RMSD between the simulated conformations and the experimental PDB structure of that peptide. However, for

Editor: Diana Murray, Weill Medical College of Cornell University, United States of America

Received December 22, 2005; **Accepted** February 20, 2005; **Published** April 14, 2006

DOI: 10.1371/journal.pcbi.0020027

Abbreviations: GB/SA, generalized-Born solvent-accessible electrostatic approximation to water; PDB, protein databank; RMSD, root-mean-square deviation

* To whom correspondence should be addressed. E-mail: bosco@maxwell.ucsf.edu

Synopsis

To carry out specific biochemical reactions, proteins must adopt precise three-dimensional conformations. During the folding of a protein, the protein picks out the right conformation out of billions of other conformations. It is not yet possible to do this computationally. Picking out the native conformation using physics-based atomically detailed models, sampled by molecular dynamics, is presently beyond the reach of computer methods. How can we speed up computational protein-structure prediction? One idea is that proteins start folding at specific parts of a chain that kink up early in the folding process. If we can identify these kinks, we should be able to speed up protein-structure prediction. Previous studies have identified likely kinks through bioinformatic analysis of existing protein structures. The goal of the authors here is to identify these putative folding initiation sites with a physical model instead. In this study, Ho and Dill show that, by chopping a protein chain into peptide pieces, then simulating the pieces in molecular dynamics, they can identify those peptide fragments that have conformational biases. These peptides identify the kinks in the protein chain.

prediction purposes—determining whether a peptide has a converged structure in the absence of knowledge of its native structure—we develop another measure based on the *backbone mesostring*, which is a coarse-grained description of the backbone conformational ensemble.

A mesostring is a one-dimensional list of the mesostates of each residue in a peptide. A mesostate refers to a discrete region of the ϕ–ψ angles of the backbone of a residue. Mesostate [a] corresponds to a helical conformation, including the α helix, 3_{10} helix, or π helix. Mesostate [b] corresponds to an extended β-strand conformation. Mesostate [l] corresponds to a left-handed helical conformation.

We use the mesostrings to cluster conformations in our simulations. Based on the three mesostates described above, an 8-mer has $3^8 = 6{,}561$ possible mesostrings. When each simulation is completed, each 8-mer peptide will have different populations for the 6,561 mesostrings, hence different free energies. The mesostring that represents the highest population (the lowest free energy) is called the *ground mesostring*. We use the properties of the ground mesostring to determine structural bias in a peptide. The ground mesostrings are classified in terms of either a reverse-turn or a helical-turn conformation (see Figure 1). We define a helical-turn as a mesostring that contains at least four [a] mesostates in a row, and a reverse-turn as a mesostring that contains either the [bab] or [baab] motifs.

How do we know when a simulation has converged? We calculate the backbone entropy using the Boltzmann formula $S = -k \Sigma_i p_i \ln p_i$, where p_i is the probability that the peptide is in mesostring i. The backbone entropy is calculated over a certain window in a trajectory, where the sum is made over only the mesostrings that are observed in the window. The backbone entropy S is useful for two purposes. First, it measures for a given peptide the sharpness of the distribution of probabilities of the mesostrings. The more peaked the distribution is, and thus the more favored a mesostring is, the lower is the backbone entropy. In this way, the backbone entropy indicates whether any one conformation is substantially favored over the others, for the given peptide. Second, the backbone entropy should converge at equilibrium, approaching an asymptotic value with time in the simulation.

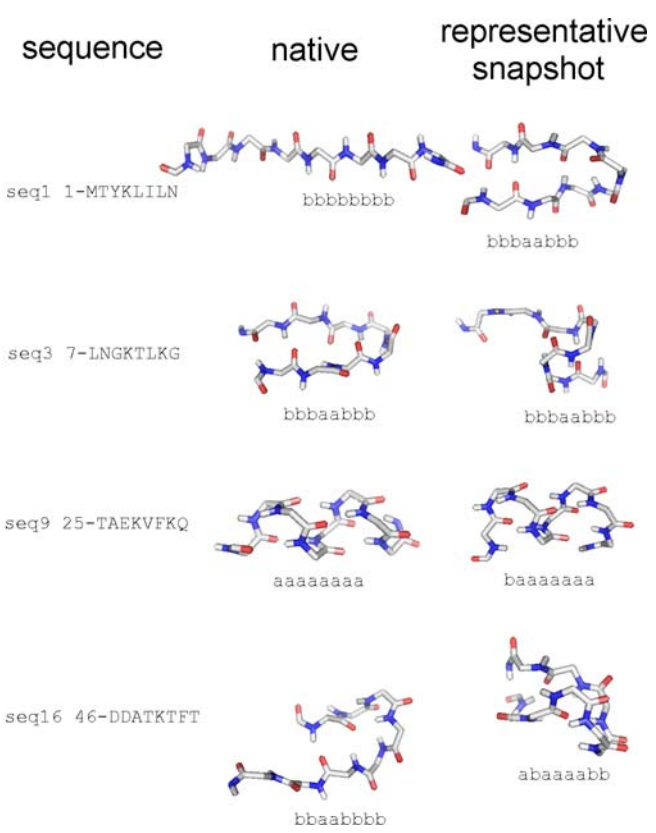

Figure 1. Representative Snapshots of Various Peptides from Protein G

The representative snapshot is the snapshot in the ground mesostring with the lowest energy. The ground mesostring of seq1 and seq3 are classed as reverse-turns, and the ground mesostring of seq9 and seq16 are classed as helical-turns.

DOI: 10.1371/journal.pcbi.0020027.g001

Even if a new mesostring emerges late within the sampling (as is often the case), it only changes the backbone entropy if it has a significant population. We use the convergence of the backbone entropy to indicate the convergence of the simulation.

We study peptide fragments extracted from a series of well-characterized proteins: protein G, protein L, protein A, and α-spectrin, and chymotrypsin inhibitor. For each peptide, we simulate the ensemble of states at equilibrium. We find that some of these peptides exhibit strong structural biases. We analyze the relationship of those structural biases to the topology of the native structure.

Results/Discussion

Structural Bias in the Peptide Conformation Ensemble

Do peptides have native-like conformations? Figure 2 shows the simulated free-energy profiles of RMSD for the peptides of protein G. We call the region of RMSD < 2 Å native-like. We find that some fragments spend a significant amount of time near their native structures (seq3, seq9, and seq10). Some peptides have a broad conformational distribution (seq14), while others have a narrow distribution (seq16). Narrow distributions indicate structural bias in the peptide. To investigate this structural bias further, we list in Table 1 the lowest free-energy mesostrings of several protein

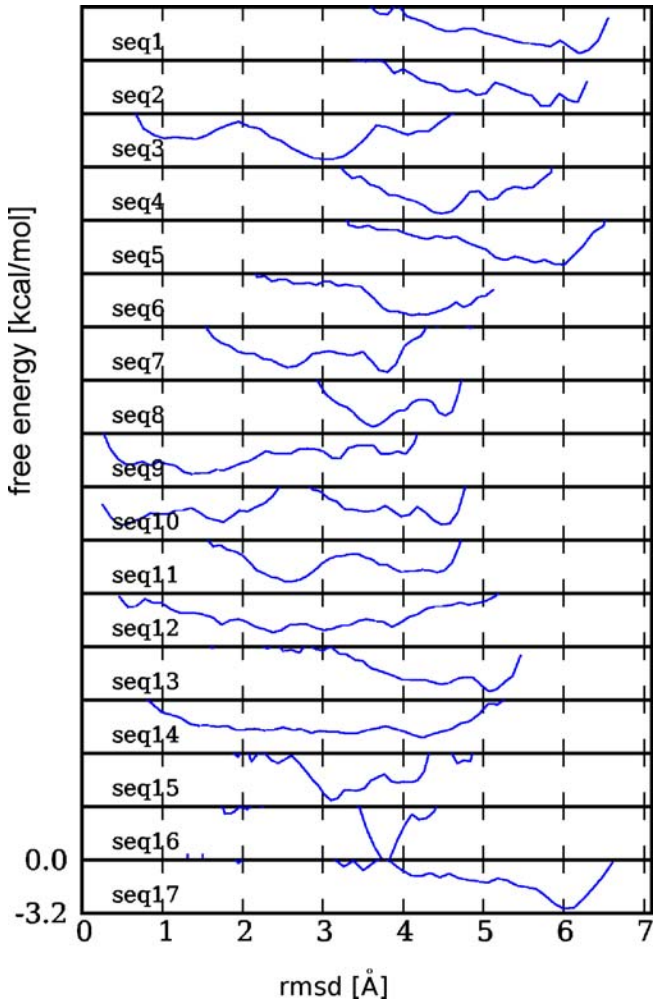

Figure 2. Free-Energy Profile of RMSD of Peptides from Protein G
DOI: 10.1371/journal.pcbi.0020027.g002

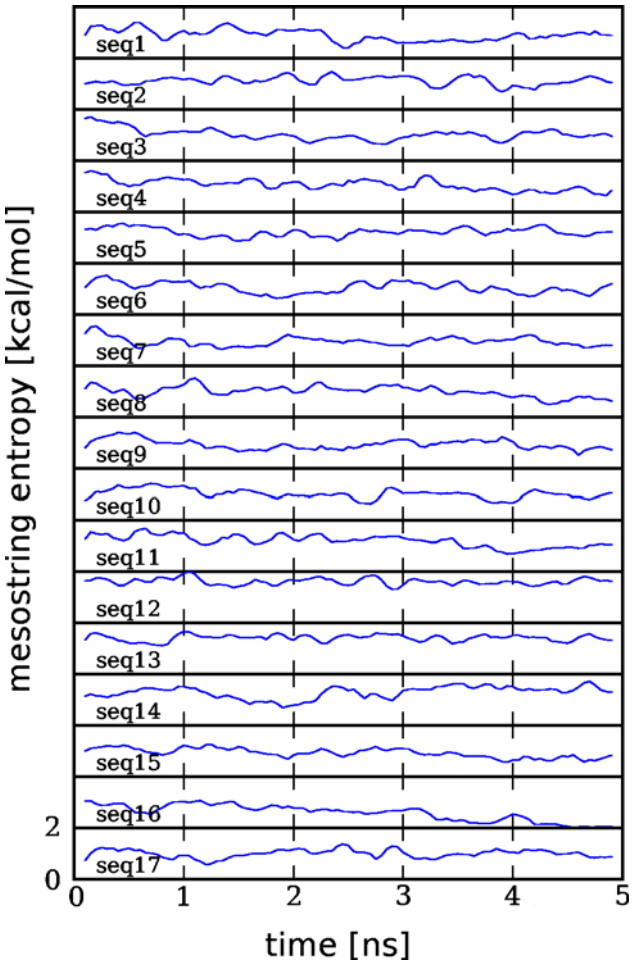

Figure 3. Variation of the Information Entropy of Mesostrings in the 17 Peptides from Protein G

The entropy at each point is calculated over a 0.2-ns window. The backbone entropy holds fairly steady after 1 ns.
DOI: 10.1371/journal.pcbi.0020027.g003

G peptides. We show in Figure 1, a representative conformation of the ground mesostrings of these peptides.

Figure 3 shows the variation in backbone entropy for the peptides of protein G. To calculate the variation in Figure 3, we deliberately chose a smaller window (0.2 ns) than the window used for the analysis (1 ns in Tables 1–4) to emphasize the fluctuations. In most of the peptides, the backbone entropy equilibrates almost immediately, with the exception of seq16, which decreases to a near zero value at about 3.5 ns. Consequently, we carry out the main analysis of the structural bias over the last 1 ns of our 5-ns trajectories. The backbone entropy specifically measures the conformation freedom in the backbone. Backbone entropy is a useful measure only when the free-energy basins in phase space are dominated by the local conformation of the backbone, and not by nonlocal interactions. As these peptides are short, nonlocal interactions should be minimal, and the backbone entropy should be the dominant entropy.

We define the existence of structural bias in a peptide in terms of two properties of the ground mesostring. First, we use the probability P_1 in observing the ground mesostring, which is derived from the relative free energies. Second, we use the free-energy gap ΔF between the ground mesostring and the next mesostring to measure the relative probability of the ground mesostring from all the other mesostrings. Specifically, we consider a peptide to have structural bias if $P_1 > 45\%$ and $\Delta F > 0.6$ kcal/mol. Of the 133 peptides we studied, we found that 48 peptides have structural bias (bold in Tables 2–4). We refer to such peptides as *structured peptides*.

Comparison of the Peptide Conformations with Native Structures

What parts of the native structure are picked out by the structured peptides? In Table 5, we list the ground mesostrings of the peptides in simulation. We highlight (in bold) the sequences that are structured and compare these structured peptides to the native secondary structures. The structured peptides adopt either a helical-turn or reverse-turn. Figure 4 shows the location of the structured peptides within the native fold topology. Below we describe the relationship between the structured peptides, the native structure, and experimental studies of the folding of these proteins.

In the protein G fragments, we find eight structured peptides that adopt a stable helical-turn conformation (Table

Table 1. Mesostrings of Various Peptides from Protein G

Peptide	Mesostring	Free Energy of the Mesostring in kcal/mol	P in Percent
seq1: 1-MTYKLILN	bbbaabbb	−3.09	31
	babaaaaa	−2.59	12
	baaaabba	−2.54	11
	baaaabbb	−2.51	11
	babaabbb	−2.31	7
	babaabba	−2.26	7
seq3: 7-LNGKTLKG	bbbaabbb	−3.29	44
	bblaabbb	−2.77	16
	abbaabbb	−2.50	10
	bbbaabbl	−2.03	4
	ablaabbb	−1.88	3
seq9: 25-TAEKVFKQ	baaaaaaa	−3.16	39
	aaaaaaaa	−3.00	26
	abaaaaaa	−2.53	11
	bbaaaaaa	−2.09	5
	aaaaaaab	−1.97	4
	bbblaaaa	−1.87	3
seq16: 46-DDATKTFT	abaaaabb	−3.69	92
	abaaaaba	−1.71	2
	aaaaabbb	−1.42	1
	baaaaaaa	−1.39	1

DOI: 10.1371/journal.pcbi.0020027.t001

2). Three of these helical-turns pick out the lone α helix in the native structure, another helical-turn picks out the turn between the helix and N-terminal β hairpin, and the remaining two helical-turns pick out the turn in the C-terminal β hairpin. Another two structured peptides with overlapping sequences adopt a stable reverse-turn conformation, which both pick out the same N-terminal hairpin-turn in the native structure. The isolated C-terminal β hairpin has been found experimentally to be stable [10], where this stability is reflected in the structural bias found in the peptide fragments of the hairpin-turn. The structured peptides provide an explanation for an ingenious study of secondary structure in protein G [39]. In that experiment, Minor and Kim replaced the α-helix sequence with a sequence based on the C-terminal hairpin. The mutant was able to fold into the same topology, showing that there is a helical propensity in the C-terminal hairpin. In the peptide studies, we find helical-turns in both the α helix and the turn of the C-terminal hairpin, which demonstrates the interchangeability of these two sequences in our simulations.

In the protein L fragments, we find four structured peptides that adopt a stable helical-turn conformation (Table 2). Two of the helical-turns pick out the α helix, while the other two helical-turns pick out the two hairpin-turns in the native structure. Another structured peptide adopts a reverse-turn conformation, which picks out the C-cap of the α helix.

In the fragments of the B domain of protein A, we found three structured peptides that adopt a stable helical-turn conformation (Table 3). These helical-turns pick out helix II and helix III, and the turn between these helices. The stability of these pieces is consistent with experimental studies of protein A fragments, which show that helix II and helix III form a stable intermediate [40].

In the myoglobin fragments, 13 structured peptides adopt a stable helical-turn conformation (Table 4). These helical-turns pick out six of the eight α helices in the native structure—with particularly long helical-turns in helices A, G, and H. Another three structured peptides adopt a stable reverse-turn conformation. Two of the reverse-turns pick out the same turn between helices G–H. The large amount of structural bias found in the fragments of helices G and H is consistent with experimental studies, which show that helices G and H form a stable intermediate [41]. Experimentally, helix F has the weakest helical propensity, and correspondingly we do not find any structured peptides in fragments of helix F.

In the chymotrypsin inhibitor fragments, we found eight structured peptides that adopt a stable helical-turn conformation (Table 4). One helical-turn picks out the 3_{10} helix in the native structure, two helical-turns pick out the α helix, one helical-turn picks out a diverging turn, and one helical-turn picks out the turn in the β hairpin. Two helical-turns erroneously pick out β strands. We also found a structured peptide that adopts a reverse-turn conformation. This reverse-turn picks out the bulge in a β strand. Experimental studies find that only the α helix is stable [42].

In the α-spectrin fragments, there are eight structured peptides that adopt stable helical-turn conformations. Two of the helical-turns erroneously pick out the RT loop. The conformation of the RT loop is somewhat indeterminate as both experimental and simulation studies (unpublished data) show that the RT loop is unstable. Another helical-turn overlaps with a diverging β-turn in the native structure. Three helical-turns erroneously pick out a β strand. The other two helical-turns pick out the turns of the two β hairpins. Experimental studies find that only a fragment of the last β hairpins has structure in solution [43].

Overall, of the 41 structured peptides that adopt a stable helical-turn conformation, 21 pick out α helices, three pick out 3_{10} helices, and two overlap with diverging turns. Because helical motifs can be considered a continuum from diverging β-turns, to 3_{10} helices, to α helices [44,45], we conclude that 26 of the helical-turns pick out helical motifs in the native structures. Another seven helical-turns pick out β-hairpin–turns, and one helical-turn is found in a helix hairpin-turn. Five helical-turns erroneously pick out β strands and two other helical-turns erroneously pick out the RT loop.

We find six structured peptides that adopt a reverse-turn conformation: one is found at a hairpin turn, two are found at strand–helix turns, three are found at helix–helix turns, and one is found at a β-strand bulge.

There is some debate [46] over whether β hairpins fold via the turn [47] or through hydrophobic clustering [48]. The results here suggest that structural bias at the turn is very important. We find that all seven β hairpins in the six proteins contain a fragment in the turn that results in a structured peptide. If we interpret the structural bias in the peptide as a kink in the full chain, then the formation of structure can be regarded as contacts coalescing around a kinky chain. In terms of the β hairpin, this does not necessarily mean that the turn forms first but that a kink favors the formation of nearby contacts.

In summary, of the 48 structured peptides found in the simulations, only five differ significantly from the native structure. Given that there are 436 residues in our six

Table 2. Ground Mesostrings of Protein G and Protein L

Protein	Peptide	Sequence	RMSD in Å	Mesostring		P_1 in Percent	ΔF in kcal/mol	TS in kcal/mol	Native Structure
				Native	Ground				
Protein G	seq1	1-MTYKLILN	5.8	bbbbbbba	bbbaabbb	31%	0.50	1.27	
	seq2	4-KLILNGKT	5.7	bbbbabaa	Bababba-	41%	0.53	1.37	
	seq3	**7-LNGKTLKG**	**3.0**	**babaabbb**	**bb-aabbb**	**60%**	**0.97**	**1.20**	**hairpin-turn**
	seq4	**10-KTLKGETT**	**4.5**	**aabbbbbb**	**aaaa-baa**	**70%**	**1.28**	**1.03**	**turn-strand**
	seq5	13-KGETTTEA	6.0	bbbbbbbb	bbbaaaab	18%	0.35	1.57	
	seq6	**16-TTTEAVDA**	**4.1**	**bbbbbaba**	**a-aaaaaa**	**48%**	**0.67**	**1.33**	**turn**
	seq7	19-EAVDAATA	3.8	bbabaaaa	abaaaabb	37%	0.15	1.19	
	seq8	**22-DAATAEKV**	**3.6**	**baaaaaaa**	**bbbbaaa-**	**64%**	**0.79**	**1.00**	**helix**
	seq9	**25-TAEKVFKQ**	**1.4**	**aaaaaaaa**	**-aaaaaaa**	**61%**	**0.93**	**1.11**	**helix**
	seq10	28-KVFKQYAN	0.5	aaaaaaaa	-aaaaaaa	40%	0.76	1.39	
	seq11	**31-KQYANDNG**	**2.5**	**aaaaaaal**	**bbaaaaa-**	**58%**	**1.02**	**1.19**	**helix-cap**
	seq12	34-ANDNGVDG	2.4	aaaalbba	bbaabbab	19%	0.65	2.08	
	seq13	37-NGVDGEWT	5.1	albbabbb	bbaa-abb	37%	1.08	1.74	
	seq14	40-DGEWTYDD	4.2	babbbbba	bbabbbbb	10%	0.11	1.93	
	seq15	**43-WTYDDATK**	**3.1**	**bbbbaaal**	**ba-aaaaa**	**64%**	**1.19**	**1.01**	**strand-turn**
	seq16	**46-DDATKTFT**	**3.8**	**baaalbbb**	**abaaaab-**	**94%**	**2.28**	**0.24**	**hairpin-turn**
	seq17	49-TKTFTVTE	6.0	albbbbba	bbaaabbb	23%	0.01	1.25	
Protein L	seq1	1-KANLIFAN	4.3	bbbbbbaa	babaaaab	42%	0.20	1.01	
	seq2	**4-LIFANGST**	**2.5**	**bbbaalbb**	**-baaaabb**	**55%**	**0.83**	**1.19**	**hairpin-turn**
	seq3	7-ANGSTQTA	5.0	aalbbbbb	babaaaab	31%	0.44	1.46	
	seq4	10-STQTAEFK	6.7	bbbbbbbb	bbaaaabb	45%	0.51	1.11	
	seq5	13-TAEFKGTF	5.7	bbbbbbba	abaaa-bb	37%	0.99	1.73	
	seq6	16-FKGTFEKA	2.5	bbbbaaaa	bbaababb	14%	0.01	1.54	
	seq7	19-TFEKATSE	3.8	baaaaaaa	babaaaab	39%	0.23	1.06	
	seq8	**22-KATSEAYA**	**4.6**	**aaaaaaaa**	**abaaaabb**	**51%**	**0.67**	**1.03**	**cap-helix**
	seq9	25-SEAYAYAD	3.8	aaaaaaaa	aaaaaaab	22%	0.13	1.44	
	seq10	28-YAYADTLK	4.2	aaaaaaab	bbbbaabb	19%	0.11	1.42	
	seq11	**31-ADTLKKDN**	**2.4**	**aaaabala**	**baaaaaa-**	**82%**	**1.16**	**0.56**	**helix-cap**
	seq12	**34-LKKDNGEY**	**2.4**	**abalallb**	**bbba-bbb**	**68%**	**1.28**	**0.96**	**turn**
	seq13	37-DNGEYTVD	5.5	lallbbbb	bbaabbab	30%	0.23	1.53	
	seq14	40-EYTVDVAD	4.4	lbbbbbbl	baaaaaaa	44%	0.78	1.23	
	seq15	**43-VDVADKGY**	**3.8**	**bbbblla**	**bbaaaa-b**	**51%**	**0.97**	**1.25**	**hairpin-turn**
	seq16	46-ADKGYTLN	3.2	blllabbb	-bbbaabb	31%	0.61	1.59	
	seq17	49-GYTLNIKFAG	6.9	labbbbbbab	bbabaaabb	15%	0.18	2.03	

RMSD is the most likely value of RMSD extracted from the free-energy profile of RMSD. The ground mesostring is sometimes nearly identical to less-populated mesostrings. If the most populated mesostrings differ by only one mesostate, we group them into a consensus mesostring, which contains one indefinite mesostate signified by [–].

P_1 is the probability of the ground mesostring.

ΔF is the free-energy difference between the ground mesostring and the next mesostring.

TS is the entropy of the mesostrings.

Native Structure is the description of the structure of the peptide in the native structure.

Bolded lines highlight structured peptides: $P_1 > 45\%$, and $\Delta F > 0.6$ kcal/mol.

DOI: 10.1371/journal.pcbi.0020027.t002

proteins, there is, on average, a kink (secondary structural indicator) approximately every nine residues along the chain.

Comparison with the I-Sites Library

Do the structural biases that are found in our simulations correlate with those in the PDB? We focus on the I-sites server (http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php), a fragment database that predicts the structures of short protein sequences [19]. In that database, predictions that have a high confidence score (>0.8) are found to predict a structure that is <1.4 Å from the native structure with a 74% probability. I-sites make eight such high-confidence predictions over four of the proteins in our dataset. Table 5 shows those successes of I-sites. Our structured peptides overlap with the I-sites predictions in six of the eight I-sites predictions. This suggests that the I-sites sequence-structure correlations are at least partly encoded in the local structural biases found in the structured peptides.

Conclusion

In this study, we have applied replica-exchange molecular dynamics, using the parm96 force field with a GB/SA solvent model, to the simulation of 133 peptide 8-mer fragments, extracted from six proteins with five different folds. We found that 48 of these peptides are strongly structured. The remaining 85 peptides have no preferred structure. Of the 48 that are structured, 41 of them fold into approximately their native conformations. In seven instances, the simulated structures are significantly inconsistent with their native structures.

Why are only 35% of the peptides structured? The reason is that by using very short peptides, we have eliminated most of the nonlocal interactions—hydrophobic clustering, cooperative helical hydrogen bonds. We thus attribute any structural bias to sidechain interactions, which will depend on specific sequence motifs.

As with all molecular dynamics simulations, the results will

Table 3. Ground Mesostrings of α-Helical Proteins

Protein	Peptide	Sequence	RMSD in Å	Mesostring		P_1 in Percent	ΔF in kcal/mol	TS in kcal/mol	Native Structure
				Native	Ground				
Protein A	seq1	1-QQNAFYEI	3.7	aaaaaaaa	baaaabbb	33	0.16	1.08	
	seq2	4-AFYEILHL	3.9	aaaaaaab	aaaaaaaa	28	0.15	1.11	
	seq3	7-EILHLPNL	3.3	aaaabaab	baaabbaa	37	0.34	1.10	
	seq4	10-HLPNLNEE	3.9	abaabbaa	abaaaabb	30	0.33	1.21	
	seq5	13-NLNEEQRN	2.9	abbaaaaa	baaaaaaa	29	0.28	1.14	
	seq6	**16-EEQRNGFI**	**4.0**	**aaaaaaaa**	**abaaa-bb**	**80**	**1.66**	**0.96**	**helix**
	seq7	19-RNGFIQSL	3.5	aaaaaaaa	ab-aabbb	19	0.53	1.83	
	seq8	22-FIQSLKDD	3.7	aaaaaaab	baaaaaaa	30	0.24	1.24	
	seq9	25-SLKDDPSQ	3.7	aaaabaaa	babaabbb	20	0.18	1.62	
	seq10	**28-DDPSQSAN**	**1.6**	**abaaaaaa**	**-baaaaaa**	**68**	**1.35**	**0.92**	**cap-helix**
	seq11	31-SQSANLLA	3.8	aaaaaaaa	babaaaab	33	0.06	1.15	
	seq12	34-ANLLAEAK	3.5	aaaaaaaa	bbaaaaaa	17	0.33	1.63	
	seq13	**37-LAEAKKLNDA**	**1.9**	**aaaaaaaaaa**	**bbaaaaaaa-**	**82**	**1.28**	**0.60**	**helix**
mvoglobin	seq1	1-MVLSEGEW	3.9	bbbbaaaa	bb-aabbb	41	0.98	1.53	
	seq2	4-SEGEWQLV	3.9	baaaaaaa	-bbaaaaa	43	0.88	1.41	
	seq3	**7-EWQLVLHV**	**1.7**	**aaaaaaaa**	**aaaaaaa-**	**83**	**1.58**	**0.76**	**helix**
	seq4	**10-LVLHVWAK**	**0.6**	**aaaaaaaa**	**-aaaaaaa**	**47**	**0.78**	**1.22**	**helix**
	seq5	13-HVWAKVEA	4.1	aaaaaaaa	bbaabbbb	26	0.13	1.37	
	seq6	16-AKVEADVA	4.0	aaaaabaa	baaabbbb	20	0.05	1.60	
	seq7	19-EADVAGHG	4.1	aabaaaaa	baaaabbb	26	0.31	1.71	
	seq8	22-VAGHGQDI	3.4	aaaaaaaa	bbbb-aab	38	0.57	1.62	
	seq9	**25-HGQDILIR**	**3.4**	**aaaaaaaa**	**bb-aaaaa**	**62**	**1.31**	**1.17**	**helix**
	seq10	28-DILIRLFK	4.5	aaaaaaaa	aaaaaaaa	49	0.51	0.88	
	seq11	**31-IRLFKSHP**	**4.1**	**aaaaaaba**	**b-aaaabb**	**76**	**1.06**	**0.81**	**helix**
	seq12	34-FKSHPETL	2.0	aaabaaaa	baabbaaa	30	0.12	1.23	
	seq13	**37-HPETLEKF**	**1.4**	**baaaaaab**	**baaaaaaa**	**67**	**1.11**	**0.76**	**helix**
	seq14	**40-TLEKFDRF**	**3.1**	**aaaabaaa**	**-aaaaaaa**	**62**	**0.66**	**0.93**	**helix**
	seq15	43-KFDRFKHL	3.5	abaaaaab	bbblabbb	31	0.13	0.96	
	seq16	46-RFKHLKTE	4.3	aaaababa	bbaaaaab	32	0.28	1.06	
	seq17	**49-HLKTEAEM**	**3.1**	**ababaaaa**	**bb-aabbb**	**56**	**1.07**	**1.18**	**turn**
	seq18	52-TEAEMKAS	2.0	baaaaaab	abaaaaaa	21	0.16	1.41	
	seq19	55-EMKASEDL	3.0	aaaabaaa	abaaabb-	30	0.49	1.55	
	seq20	58-ASEDLKKA	1.9	abaaaaaa	aaaaaaaa	37	0.28	1.10	
	seq21	**61-DLKKAGVT**	**3.9**	**aaaaaaaa**	**baaaa-bb**	**59**	**1.00**	**1.30**	**helix**
	seq22	64-KAGVTVLT	4.3	aaaaaaaa	babaabab	19	0.11	1.47	
	seq23	67-VTVLTALG	3.7	aaaaaaaa	aaaaaaa-	40	0.75	1.48	
	seq24	70-LTALGAIL	4.1	aaaaaaaa	baab-aab	31	0.69	1.80	
	seq25	73-LGAILKKK	3.5	aaaaaaal	bba-aaaa	32	0.70	1.62	
	seq26	76-ILKKKGHH	3.9	aaaallba	bbaaaabb	28	0.57	1.44	
	seq27	79-KKGHHEAE	3.5	allbaaaa	ab-aabbb	23	0.54	1.66	
	seq28	82-HHEAELKP	4.7	baaaaaaa	aaaaaabb	37	0.04	0.91	
	seq29	85-AELKPLAQ	3.8	aaaaaaaa	bbbbbbbb	13	0.15	1.76	
	seq30	88-KPLAQSHA	4.3	aaaaaaaa	bbbaabbb	29	0.25	1.29	
	seq31	91-AQSHATKH	3.3	aaaaaaaa	babaaaab	29	0.44	1.40	
	seq32	94-HATKHKIP	2.5	aaaaalbb	baaabbbb	25	0.30	1.34	
	seq33	97-KHKIPIKY	3.5	aalbbaaa	aaabaaa-	40	0.61	1.39	
	seq34	**100-IPIKYLEF**	**3.4**	**bbaaaaaa**	**bbbaaabb**	**58**	**0.90**	**0.95**	**helix**
	seq35	103-KYLEFISE	4.2	aaaaaaaa	ba-aabbb	45	0.53	1.25	
	seq36	106-EFISEAII	4.6	aaaaaaaa	babaaaab	35	0.13	0.99	
	seq37	**109-SEAIIHVL**	**0.6**	**aaaaaaaa**	**b-aaaaaa**	**73**	**1.12**	**0.90**	**helix**
	seq38	**112-IIHVLHSR**	**4**	**aaaaaaaa**	**aaaaaabb**	**48**	**0.72**	**0.94**	**helix**
	seq39	**115-VLHSRHPG**	**3.7**	**aaaaabaa**	**bbbaabbb**	**49**	**0.81**	**1.14**	**turn**
	seq40	**118-SRHPGNFG**	**3.1**	**aabaaabb**	**bbbbba-b**	**47**	**0.98**	**1.48**	**turn**
	seq41	121-PGNFGADA	3.9	aaabbaaa	bbbbb-bb	10	0.43	2.31	
	seq42	124-FGADAQGA	3.6	bbaaaaaa	bbaabb-b	18	0.72	2.14	
	seq43	127-DAQGAMNK	4.3	aaaaaaaa	bbbbabb-	35	0.80	1.66	
	seq44	130-GAMNKALE	4.1	aaaaaaaa	bbbaabbb	22	0.26	1.50	
	seq45	133-NKALELFR	3.7	aaaaaaaa	baaabaab	36	0.28	1.13	
	seq46	**136-LELFRKDI**	**0.4**	**aaaaaaaa**	**-aaaaaaa**	**78**	**0.88**	**0.73**	**helix**
	seq47	**139-FRKDIAAK**	**0.4**	**aaaaaaaa**	**b-aaaaaa**	**71**	**1.28**	**0.95**	**helix**
	seq48	**142-DIAAKYKE**	**3.9**	**aaaaaaaa**	**aaaaaaaa**	**46**	**0.68**	**1.08**	**helix**
	seq49	145-AKYKELGYQG	3.7	aaaaaalala	babaaaabb-	38	0.81	1.66	

RMSD is the most likely value of RMSD extracted from the free-energy profile of RMSD. The ground mesostring is sometimes nearly identical to less-populated mesostrings. If the most populated mesostrings differ by only one mesostate, we group them into a consensus mesostring, which contains one indefinite mesostate signified by [–].
P_1 is the probability of the ground mesostring.

ΔF is the free-energy difference between the ground mesostring and the next mesostring.
TS is the entropy of the mesostrings.
Native Structure is the description of the structure of the peptide in the native structure.
Bolded lines highlight structured peptides: $P_1 > 45\%$, and $\Delta F > 0.6$ kcal/mol.
DOI: 10.1371/journal.pcbi.0020027.t003

depend somewhat on the choice of force field. One limitation, for example, is that none of the current force fields model the backbone very well, especially in glycine. Neither can current force fields model the left-handed α-helical conformation accurately, resulting in the paucity of ground mesostrings containing the [l] mesostate. Better force fields may improve our predictions. As we simulated only short peptides, we have eliminated various cooperative nonlocal interactions—interactions that are particularly sensitive to specific details of the force field.

The I-sites library taken from PDB peptide preferences makes eight high-confidence predictions in four of the six proteins. In those instances, our simulations are largely consistent with theirs, indicating that the intrinsic physical preferences contribute to the PDB structures. However, the present simulations are also more informative, giving 48 structures (with 85% reliability) among the 133 peptides we tested, in contrast to the eight (having 74% reliability) found by I-sites.

Current structure-prediction systems rely on a pragmatic mix of bio-informatics and physical modeling [23,24]. A key component of these systems is the use of fragment libraries to

Table 4. Ground Mesostrings of β-Sheet Proteins

Protein	Peptide	Sequence	RMSD in Å	Mesostring		P_1 in Percent	ΔF in kcal/mol	TS in kcal/mol	Native Structure
				Native	Ground				
Chymotrypsin inhibitor	seq1	1-NLKTEWPE	5.2	bbbabbaa	Bbaaabb-	65	0.89	0.90	**loop**
	seq2	4-TEWPELVG	4.2	abbaaabl	b-baaaab	85	1.6	0.57	**3_{10} helix**
	seq3	7-PELVGKSV	2.9	aaablbba	baabbaab	15	0.24	1.65	
	seq4	10-VGKSVEEA	4.1	blbbaaaa	abaaaabb	34	0.51	1.39	
	seq5	13-SVEEAKKV	0.5	baaaaaaa	-aaaaaaa	63	0.82	0.92	**helix**
	seq6	16-EAKKVILQ	4.3	aaaaaaaa	baaaaaaa	24	0.14	1.32	
	seq7	19-KVILQDKP	3.9	aaaaaaba	babaaabb	34	0.45	1.16	
	seq8	22-LQDKPEAQ	2.7	aaabaabb	Bbabaaa-	45	0.81	1.39	**helix-cap**
	seq9	25-KPEAQIIV	4.9	baabbbbb	bbaaabbb	41	0.71	1.17	
	seq10	28-AQIIVLPV	5.7	bbbbbbbb	b-aaabbb	64	0.68	0.91	**strand**
	seq11	31-IVLPVGTI	3.1	bbbbblbb	bbbaaaab	20	0.14	1.54	
	seq12	34-PVGTIVTM	4.2	bblbbbba	bbbaaaaa	12	0.06	1.87	
	seq13	37-TIVTMEYR	4.0	bbbbabbb	aaaaabba	30	0.44	1.23	
	seq14	40-TMEYRIDR	3.7	babbbaab	bb-aaaaa	64	0.87	1.02	**loop-turn**
	seq15	43-YRIDRVRL	3.2	bbaabbbb	bbbaaaaa	48	0.17	0.75	
	seq16	46-DRVRLFVD	6.4	abbbbbbb	abbaabbb	40	0.64	1.18	
	seq17	49-RLFVDKLD	4.2	bbbbbaal	babaaabb	37	0.07	0.83	
	seq18	52-VDKLDNIA	4.1	bbaalbba	ba-aaabb	64	1.06	1.09	**hairpin-turn**
	seq19	55-LDNIAEVP	3.3	albbabbb	babaaabb	22	0.15	1.36	
	seq20	58-IAEVPRVG	3.7	babbbbbb	baabba-b	66	0.97	0.99	bulge
α Spectrin	seq1	1-KELVLALY	4.3	bbbbbbab	-aaaaaaa	64	0.90	1.01	**strand**
	seq2	4-VLALYDYQ	3.7	bbbabbbb	aaaaaaaa	34	0.31	1.29	
	seq3	7-LYDYQEKS	4.0	abbbbbab	baaaaaaa	55	0.78	0.88	**loop**
	seq4	10-YQEKSPRE	3.6	bbbabaab	baaabbaa	44	0.53	0.90	
	seq5	13-KSPREVTM	3.8	abaabbbb	Bbbbaaa-	58	1.01	1.06	**loop**
	seq6	16-REVTMKKG	4.5	abbbbbbl	abaaaaa-	48	0.86	1.18	**diverging-turn**
	seq7	19-TMKKGDIL	2.7	bbbblbbb	babbbbbb	23	0.18	1.45	
	seq8	22-KGDILTLL	4.4	blbbbbba	b-baaaaa	78	1.73	0.97	**strand**
	seq9	25-ILTLLNST	3.9	bbbbabaa	b-aaaaaa	75	1.42	0.93	**strand**
	seq10	28-LLNSTNKD	4.0	babaabaa	Bbbaaaa-	53	1.06	1.25	**hairpin-turn**
	seq11	31-STNKDWWK	3.2	aabaabbb	bbbbaabb	36	0.22	1.15	
	seq12	34-KDWWKVEV	5.8	aabbbbbb	b-baabbb	43	0.70	1.32	
	seq13	37-WKVEVNDR	3.8	bbbbblab	bbbaaaaa	46	0.60	1.01	**hairpin-turn**
	seq14	40-EVNDRQGF	3.7	bblabbbb	baaaaabb	25	0.07	1.33	
	seq15	43-DRQGFVPA	5.6	abbbbbba	abbbabbb	12	0.07	1.62	
	seq16	46-GFVPAAYV	3.2	bbbbaaab	bbbaaabb	36	0.78	1.42	
	seq17	49-PAAYVKKLD	3.3	baaabbbbb	abaaaaaaa	41	0.14	0.93	

RMSD is the most likely value of RMSD extracted from the free-energy profile of RMSD. The ground mesostring is sometimes nearly identical to less-populated mesostrings. If the most populated mesostrings differ by only one mesostate, we group them into a consensus mesostring, which contains one indefinite mesostate signified by [–].
P_1 is the probability of the ground mesostring.
ΔF is the free-energy difference between the ground mesostring and the next mesostring.
TS is the entropy of the mesostrings.
Native Structure is the description of the structure of the peptide in the native structure.
DOI: 10.1371/journal.pcbi.0020027.t004

Table 5. Comparison of the Structural Bias with the Native Structure

Protein	Lines	Structure per Residue
Protein G	1	MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
	2	-SSSSSSS----SSSSSSSS--HHHHHHHHHHHHH------SSSSSTT---SSSSS--
	3	bbbbbbbabaabbbbbbbbbbabaaaaaaaaaaaaaaaaalbbabbbbbbaaalbbbbba
	4	---------aa------aaaaaa--aaaaaaaaaaaa--------aaaaaa-----
	5	__HHHHHHHHHHHHH S_TTT__S
Protein L	1	KANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEYTVDVADKGYTLNIKFAG
	2	-SSS-------SSS------HHHHHHHHHHHH--------SS---S----S---SS--
	3	bbbbbbaalbbbbbbbbbbaaaaaaaaaaaaaaaabalallbbbbbbllabbbbbbab
	4	-----aaaa-----aaa------aaaa----aaaaar-------aaaa----------
Protein A	1	QQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDA
	2	-HHHHHHHH------HHHHHHHHHHHH-TT-HHHHHHHHHHHH-
	3	aaaaaaaaaabaabbaaaaaaaaaaaaabaaaaaaaaaaaaaaaa
	4	----------------aaa---------aaaaaaa--aaaaaaa-
	5	__HHHHHHHHH HH__GGG
Myoglobin	1	MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKAGVTVLT
	2	----HHHHHHHHHHHHHHH--HHHHHHHHHHHHH--HHH33-TT-TT---HHHHHH--HHHHHHHHHHH
	3	bbbbaaaaaaaaaaaaaaabaaaaaaaaaaaaaaabaaaaaabaaaaabababaaaaaabaaaaaaaaaaa
	4	------aaaaaaaaaa----------aaaaaaaaaa-aaaaaaaaaa----rr--------aaaa------
	5	HHHH_GGG
	1	ALGAILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKD
	2	HHHHHHH------HHHHHHHHHHHHH-----HHHHHHHHHHHHHHHHH-TT---HHHHHHHHHHHHHHHH
	3	aaaaaaaallbaaaaaaaaaaaaaaalbbaaaaaaaaaaaaaaaaaabaaabbaaaaaaaaaaaaaaaa
	4	--------------------------------aaaa----aaaaaaarr---r-------------aaaaaa
	5	___HHHH HHHHHHHHHH
	1	IAAKYKELGYQG
	2	HHHHHHHH
	3	aaaaaaaalala
	4	aaaa--------
Chymotrypsin inhibitor 2	1	NLKTEWPELVGKSVEEAKKVILQDKPEAQIIVLPVGTIVTMEYRIDRVRLFVDKLDNIAEVPRVG
	2	----S-333TT---HHHHHHHHH-TT-SSSSS----------TTSSSSSS-TT--S----S--
	3	bbbabbaaablbbaaaaaaaaaaabaabbbbbbbblbbbbabbbaabbbbbbbaalbbabbbbbb
	4	--aaa-aaaa---aaaaaaa-----aaa-aaa----------aaaaa-------aaa-rr-----
α Spectrin	1	KELVLALYDYQEKSPREVTMKKGDILTLLNSTNKDWWKVEVNDRQGFVPAAYVKKLD
	2	--SSSS---S----TT---S-TT-SSSSSS-------SSS-TT-SSSSS333SSSS-
	3	bbbbbbabbbbbabaabbbbbbblbbbbbabaabaabbbbbbblabbbbbbaaabbbbb
	4	-aaaaaaaaaaaaaa-aaaaaa--aaaaaaaaaa-----aaaaa----------
	5	EETT_E

Line 1 is the amino-acid sequence.
Line 2 is the secondary structure in the native structure (3, 3_{10} helix; H, α-helical; S, sheet; T, H-bonded–turn).
Line 3 is the mesostring of the native structure.
Line 4 is the ground mesostring predicted from the peptides (aaa, helical-turn; rr, reverse-turn).
Line 5 is the I-sites predictions (_, other; E, extended but not H-bonded; G, other helical-turn; H, α-helical; S, sheet; T, turn).
DOI: 10.1371/journal.pcbi.0020027.t005

identify folding initiation sites. Here we have identified the physical origin of the sequence–structure relations identified in the fragment libraries—local structural bias in short peptide sequences. The calculations are not exorbitant, as each peptide takes ~160 CPU node hours, and, in many cases, our results go beyond the fragment libraries. By replacing fragment libraries with peptide simulations to identify folding initiation sites, we move closer to the goal of predicting protein structures using only physical models.

Materials and Methods

Replica-exchange simulations of the peptides. Replica-exchange simulations were conducted using a PERL wrapper (http://www.dillgroup.ucsf.edu/~jchodera/code/rex) around the SANDER molecular dynamics program for the Amber7 molecular-modeling package [49]. We used 16 replicas exponentially spaced between 270K and 690K, achieving an exchange–acceptance probability of approximately 50%. Exchanges were attempted every 1 ps, with constant-energy dynamics conducted between exchanges. After each exchange attempt, the velocities were redrawn from the appropriate Maxwell-Boltzmann distribution to ensure proper thermostating. A 2-fs time step was used, and bonds to hydrogens were constrained with SHAKE [50]. Configurations were stored every 1 ps for analysis. Simulations were run for 5 ns per replica and the first 4 ns were used for equilibration. The peptides were capped with ACE and NME blocking groups, and initialized in the extended state. Systems were set up using the LEAP program. Peptide parameters were taken from the Amber Parm96 force field, and the GB/SA model of Tsui and Case was used [37], along with a surface area penalty term of 5 cal · mol^{-1} · $Å^{-2}$.

Calculating thermodynamic observables. We use replica exchange [36] to simulate the equilibrium ensemble. It samples k parallel replicas, each of which is at a different temperature. Hence, to extract thermodynamic observables for a given temperature, say T = 300K, we must reweigh the configurations taken from the k different temperatures β^k in order to combine them into a representative ensemble. We do this reweighing of the replicas with an implementation [51] of the Weighted Histogram Analysis Method [52].

We first calculate the dimensionless free-energy f^k for each replica k. Starting with a crude estimate of f^k, we calculate $\Omega^k{}_E$—the weight of states with energy E in replica k:

$$\Omega^{k}_{E} = \frac{N^{k}_{E}}{N^{k} \exp(f^{k} - \beta^{k} E)} \quad (1)$$

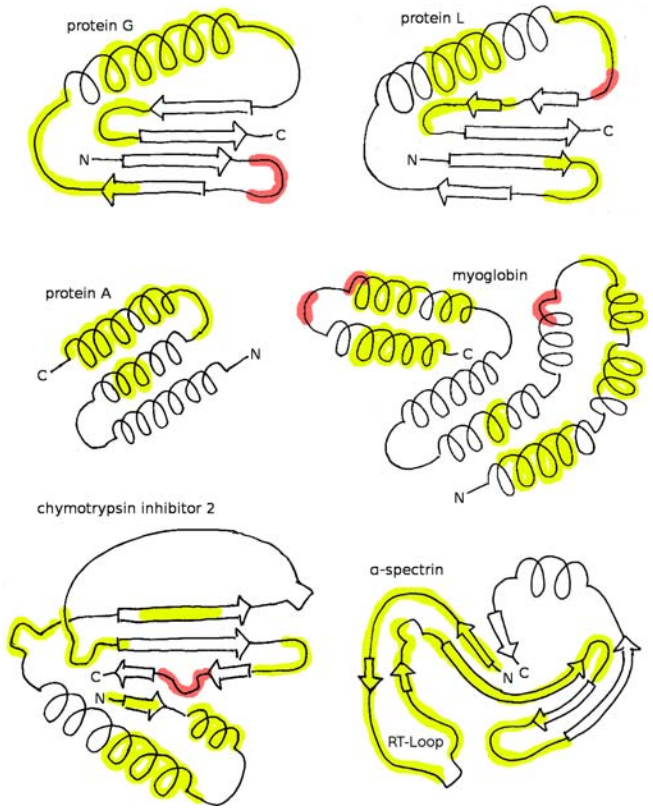

Figure 4. Distribution of Kinks in the Topology of Protein Structures

α helices and 3_{10} helices are drawn as coils, hydrogen-bonded β-turns are drawn as a notch, and β strands are drawn as arrows.

Yellow indicates helical-turn [-aaa-] and red indicates reverse-turns [baab]. 11 of the 14 α helices contain a helical-turn. The turns of all seven β hairpins contain a peptide fragment that is structured.

DOI: 10.1371/journal.pcbi.0020027.g004

where N^k_E is the number of snapshots in replica k with energy E. From the distribution of Ω^k_E, we calculate a new estimate of f^k by

$$f^k = -\log\left[\sum_E \Omega^k_E \exp(\beta^k E)\right] \quad (2)$$

We iterate the above two steps until f^k converges. Then we use these dimensionless free energies f^k to reweigh the relative free-energy profile F of observable x to the target temperature β_{tar}:

$$F_x(\beta_{tar}) = -\frac{1}{\beta_{tar}}\log\left\{\sum_E\left[\frac{\sum_k N^k_{x,E}\exp(\beta_{tar}E)}{\sum_{k'} N^{k'}_{x,E}\exp(f^{k'}-\beta^{k'}E)}\right]\right\} \quad (3)$$

After using the Weighted Histogram Analysis Method to calculate the relative free energies F_i of a mesostring i, we calculate the probabilities P_i by

$$P_i = \frac{\exp(-\beta_{tar}F_i)}{\sum_{i'}\exp(-\beta_{tar}F_{i'})} \quad (4)$$

When we merge similar mesostrings into a consensus mesostring, we calculate the free-energy difference to another mesostring j by

$$\Delta F = -\frac{1}{\beta_{tar}}\log\left(\frac{P_{consensus}}{P_j}\right) \quad (5)$$

Defining the backbone mesostates. A key part of our analysis is the discretizing of the backbone degrees of freedom. This is based on the original analysis of the protein backbone [53]. In that analysis, Ramachandran and coworkers showed that the stereochemistry of the protein backbone breaks up the backbone φ–ψ angles into three distinct regions, each separated by significant energy barriers. We can thus describe the conformation of a peptide as a string of discrete mesostates—we call this the mesostring. A given mesostring is separated in energy from other mesostrings. Each mesostring corresponds to a low-energy basin in the conformation space of the peptide backbone. It is then straightforward to extract the local structure from the lowest free-energy basin. This partitioning in terms of discrete regions in the backbone angles has been observed in a molecular dynamics simulation of an α-helical peptide [31].

The original analysis of the backbone identified three distinct regions in the φ–ψ angles [53]. Recent studies of the protein database found that these three regions can be further divided up into five clusters of density [54,55]. Some of the barriers between these five regions are small, which leaves three regions separated by large barriers. However we cannot use the database analysis to define the boundaries of the backbone mesostates because current force fields cannot replicate the database distribution of φ–ψ angles. We must define the boundaries the backbone mesostates in terms of the force field in our molecular dynamics: we ran replica-exchange simulations of the alanine dipeptide and the glycine dipeptide for 10 ns and calculated the free-energy profile of the φ–ψ angles in bins of 5°. Based on the resultant free-energy profile, we break up the Ramachandran plot in terms of the following mesostates:

$$[b] : (-180° < \varphi < 0°, 45° < \psi < 180°)$$
$$\mathrm{U}(-180° < \varphi < 0°, -180° < \psi < -135°)$$
$$\mathrm{U}(120° < \varphi < 180°, 45° < \psi < 180°)$$
$$\mathrm{U}(120° < \varphi < 180°, -180° < \psi < -135°)$$

$$[a] : (-180° < \varphi < 0°, -135° < \psi < 45°)$$
$$\mathrm{U}(120° < \varphi < 180°, -135° < \psi < 45°)$$

$$[l] : (0° < \varphi < 120°, -180° < \psi < 180°)$$
$$\mathrm{U}(120° < \varphi < 180°, -135° < \psi < 45°)$$

And for glycine:

$$[b] : (-180° < \varphi < 0°, 45° < \psi < 180°)$$
$$\mathrm{U}(-180° < \varphi < 0°, -180° < \psi < -135°)$$
$$\mathrm{U}(0° < \varphi < 180°, 135° < \psi < 180°)$$
$$\mathrm{U}(0° < \varphi < 180°, -180° < \psi < -45°)$$

$$[a] : (-180° < \varphi < 0°, -135° < \psi < 45°)$$

$$[l] : (0° < \varphi < 180°, -45° < \psi < 135°)$$

Acknowledgments

Thanks to John Chodera for the replica-exchange wrapper for the molecular dynamics package. Thanks to Banu Ozkan, Vince Voelz and Albert Wu for many invaluable discussions.

Author contributions. BKH and KAD conceived and designed the experiments. BKH performed the experiments. BKH analyzed the data. BKH wrote the paper.

Funding. We appreciate the support of NIH grant GM34993.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Dyson HJ, Wright PE (1998) Equilibrium NMR studies of unfolded and partially folded proteins. Nat Struct Biol 5 (Supplement): 499–503.
2. Dyson HJ, Merutka G, Waltho JP, Lerner RA, Wright PE (1992) Folding of peptide fragments comprising the complete sequence of proteins. Models for initiation of protein folding. I. Myohemerythrin. J Mol Biol 226: 795–817.
3. Shin HC, Merutka G, Waltho JP, Tennant LL, Dyson HJ, Wright PE (1993) Peptide models of protein folding initiation sites. 3. The G–H helical hairpin of myoglobin. Biochemistry 32: 6356–6364.
4. Waltho JP, Feher VA, Merutka G, Dyson HJ, Wright PE (1993) Peptide models of protein folding initiation sites. 1. Secondary structure formation by peptides corresponding to the G- and H-helices of myoglobin. Biochemistry 32: 6337–6347.

5. Ramirez-Alvarado M, Serrano L, Blanco FJ (1997) Conformational analysis of peptides corresponding to all the secondary structure elements of protein L B1 domain: Secondary structure propensities are not conserved in proteins with the same fold. Protein Sci 6: 162–174.
6. Eliezer D, Chung J, Dyson HJ, Wright PE (2000) Native and non-native secondary structure and dynamics in the pH 4 intermediate of apomyoglobin. Biochemistry 39: 2894–2901.
7. Mohana-Borges R, Goto NK, Kroon GJ, Dyson HJ, Wright PE (2004) Structural characterization of unfolded states of apomyoglobin using residual dipolar couplings. J Mol Biol 340: 1131–1142.
8. Marqusee S, Robbins VH, Baldwin RL (1989) Unusually stable helix formation in short alanine-based peptides. Proc Natl Acad Sci U S A 86: 5286–5290.
9. Munoz V, Serrano L (1994) Elucidating the folding problem of helical peptides using empirical parameters. Nat Struct Biol 1: 399–409.
10. Blanco FJ, Rivas G, Serrano L (1994) A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. Nat Struct Biol 1: 584–590.
11. Searle MS, Williams DH, Packman LC (1995) A short linear peptide derived from the N-terminal sequence of ubiquitin folds into a water-stable non-native beta-hairpin. Nat Struct Biol 2: 999–1006.
12. Zerella R, Evans PA, Ionides JM, Packman LC, Trotter BW, Mackay JP, Williams DH (1999) Autonomous folding of a peptide corresponding to the N-terminal beta-hairpin from ubiquitin. Protein Sci 8: 1320–1331.
13. Espinosa JF, Munoz V, Gellman SH (2001) Interplay between hydrophobic cluster and loop propensity in beta-hairpin formation. J Mol Biol 306: 397–402.
14. Rotondi KS, Gierasch LM (2003) Role of local sequence in the folding of cellular retinoic abinding protein I: Structural propensities of reverse turns. Biochemistry 42: 7976–7985.
15. Eisenberg D, Weiss RM, Terwilliger TC (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci U S A 81: 140–144.
16. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262: 1680–1685.
17. Han KF, Baker D (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc Natl Acad Sci U S A 93: 5814–5818.
18. Bystroff C, Simons KT, Han KF, Baker D (1996) Local sequence–structure correlations in proteins. Curr Opin Biotechnol 7: 417–421.
19. Bystroff C, Baker D (1998) Prediction of local structure in proteins using a library of sequence–structure motifs. J Mol Biol 281: 565–577.
20. Tsai CJ, Maizel JV Jr, Nussinov R (2000) Anatomy of protein structures: Visualizing how a one-dimensional protein chain folds into a three-dimensional shape. Proc Natl Acad Sci U S A 97: 12038–12043.
21. Kolodny R, Koehl P, Guibas L, Levitt M (2002) Small libraries of protein fragments model native protein structures accurately. J Mol Biol 323: 297–307.
22. Tendulkar AV, Joshi AA, Sohoni MA, Wangikar PP (2004) Clustering of protein structural fragments reveals modular building block approach of nature. J Mol Biol 338: 611–629.
23. Aloy P, Stark A, Hadley C, Russell RB (2003) Predictions without templates: New folds, secondary structure, and contacts in CASP5. Proteins 53 (Supplement 6):: 436–456.
24. Moult J (2005) A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15: 285–289.
25. Avbelj F, Moult J (1995) Determination of the conformation of folding initiation sites in proteins by computer simulation. Proteins 23: 129–141.
26. Srinivasan R, Rose GD (1995) LINUS: A hierarchic procedure to predict the fold of a protein. Proteins 22: 81–99.
27. Gibbs N, Clarke AR, Sessions RB (2001) Ab initio protein structure prediction using physicochemical potentials and a simplified off-lattice model. Proteins 43: 186–202.
28. Klepeis JL, Floudas CA (2002) Ab initio prediction of helical segments in polypeptides. J Comput Chem 23: 245–266.
29. Daura X, van Gunsteren WF, Mark AE (1999) Folding–unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. Proteins 34: 269–280.
30. de Groot BL, Daura X, Mark AE, Grubmuller H (2001) Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. J Mol Biol 309: 299–313.
31. Mu Y, Nguyen PH, Stock G (2005) Energy landscape of a small peptide revealed by dihedral angle principal component analysis. Proteins 58: 45–52.
32. Bystroff C, Garde S (2003) Helix propensities of short peptides: Molecular dynamics versus bioinformatics. Proteins 50: 552–562.
33. Kim PS, Baldwin RL (1982) Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu Rev Biochem 51: 459–489.
34. Dill KA, Fiebig KM, Chan HS (1993) Cooperativity in protein-folding kinetics. Proc Natl Acad Sci U S A 90: 1942–1946.
35. Baldwin RL, Rose GD (1999) Is protein folding hierarchic? I. Local structure and peptide folding. Trends Biochem Sci 24: 26–33.
36. Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters 314: 141–151.
37. Tsui V, Case DA (2000) Theory and applications of the generalized Born solvation model in macromolecular simulations. Biopolymers 56: 275–291.
38. Zhou R (2003) Free energy landscape of protein folding in water: Explicit vs. implicit solvent. Proteins 53: 148–161.
39. Minor DL Jr, Kim PS (1996) Context-dependent secondary structure formation of a designed protein sequence. Nature 380: 730–734.
40. Bai Y, Englander SW (1994) . Bai Y, Englander SW (1994) Hydrogen bond strength and beta-sheet propensities: The role of a side chain blocking effect. Proteins 18, 262–266.
41. Jennings PA, Wright PE (1993) Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. Science 262: 892–896.
42. Itzhaki LS, Neira JL, Ruiz-Sanz J, de Prat Gay G, Fersht AR (1995) Search for nucleation sites in smaller fragments of chymotrypsin inhibitor 2. J Mol Biol 254: 289–304.
43. Viguera AR, Jimenez MA, Rico M, Serrano L (1996) Conformational analysis of peptides corresponding to beta-hairpins and a beta-sheet that represent the entire sequence of the alpha-spectrin SH3 domain. J Mol Biol 255: 507–521.
44. Sundaralingam M, Sekharudu YC (1989) Water-inserted alpha-helical segments implicate reverse turns as folding intermediates. Science 244: 1333–1337.
45. Soman KV, Karimi A, Case DA (1991) Unfolding of an alpha-helix in water. Biopolymers 31: 1351–1361.
46. Du D, Zhu Y, Huang CY, Gai F (2004) Understanding the key factors that control the rate of beta-hairpin folding. Proc Natl Acad Sci U S A 101: 15915–15920.
47. Klimov DK, Thirumalai D (2000) Mechanisms and kinetics of beta-hairpin formation. Proc Natl Acad Sci U S A 97: 2544–2549.
48. Munoz V, Thompson PA, Hofrichter J, Eaton WA (1997) Folding dynamics and mechanism of beta-hairpin formation. Nature 390: 196–199.
49. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, et al. (1995) Amber, a package of computer-programs for applying molecular mechanics, normal-mode analysis, molecular-dynamics and free-energy calculations to simulate the structural and energetic properties of molecules. Comput Physics Commun 91: 1–41.
50. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical-integration of Cartesian equations of motion of a system with constraints—molecular-dynamics of N-alkanes. Journal of Computational Physics 23: 327–341.
51. Chodera JD, Swope WC, Pitera JW, Seok C, Dill KA (2006) Use of the Weighted Histogram Analysis Method for the analysis of simulated and parallel tempering simulations. J Chem Theory Comput: In press.
52. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM (1992) The Weighted Histogram Analysis Method for free-energy calculations on biomolecules. 1. The method. J Comput Chem 13: 1011–1021.
53. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. J Mol Biol 7: 95–99.
54. Karplus PA (1996) Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci 5: 1406–1420.
55. Ho BK, Thomas A, Brasseur R (2003) Revisiting the Ramachandran plot: Hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. Protein Sci 12: 2508–2522.